Molecular Epidemiology Workshop: Structured Coalescents and their Applications

Jay Taylor and Ananias Escalante

School of Mathematical and Statistical Sciences
School of Life Sciences
Center for Evolutionary Medicine and Informatics
Arizona State University
Outline

1. Structured Coalescents: Derivation
2. Structured Coalescent: Simulation and Properties
3. Structured Coalescents: Inference
4. Structured Coalescents with Selection
Coalescents and Exchangeability

We saw previously that for many models Kingman’s coalescent provides a good approximation to the genealogy of a random sample. In part, this is because these models share a property called **exchangeability**, meaning that the behavior of the model does not depend on how we label the individuals in the population.

However, exchangeability can be violated in several biologically important ways:

1. If the population is subdivided into $D$ subpopulations, then pairs of individuals occupying the same deme will not be exchangeable with pairs of individuals occupying different demes.

2. If individuals differ in fitness, then pairs of individuals with the same fitness will not be exchangeable with pairs that have different fitnesses.

3. Exchangeability can also be violated by non-random sampling. For example, a pair of individuals carrying allele $A$ will not be exchangeable with a pair containing one $A$ allele and one $B$ allele.
The problem: The difficulty created when individuals are not exchangeable is that the genealogy of a random sample is no longer a Markov process. Unfortunately, without the Markov property, analysis and simulation become prohibitively difficult.

A solution: The central idea underlying structured coalescents is that we can recover the Markov property by keeping track of certain additional information specifying the states of the ancestral lineages.
Consider the following model for a subdivided population introduced by Notohara (1990).

1. The population is subdivided into $D$ demes, having haploid population sizes $N_1, \ldots, N_D$. Both the number of demes and the number of individuals surviving to adulthood in each deme is constant.

2. Generations are non-overlapping.

3. Migration occurs at the beginning of each generation and different individuals migrate independently of one another. Specifically, each individual in deme $i$ has probability $m_{ij}$ of migrating to deme $j$.

4. Following migration, each of the $N_i + M_i$ adults located in deme $i$ gives birth to a large number of offspring. However, only $N_i$ of these offspring, chosen uniformly at random, survive to adulthood.
Notohara’s Model: \( D = 2, \ N_1 = 5 \) and \( N_2 = 3 \)
Notohara’s Model: $D = 2$, $N_1 = 5$ and $N_2 = 3$
Notohara’s Model: $D = 2$, $N_1 = 5$ and $N_2 = 3$
Notohara’s Model: \( D = 2, N_1 = 5 \) and \( N_2 = 3 \)
Notohara’s Model: \( D = 2, N_1 = 5 \) and \( N_2 = 3 \)
Notohara’s Model: \( D = 2, \ N_1 = 5 \) and \( N_2 = 3 \)
Suppose that we have sampled two individuals at random from known locations within the population.

- In this setting, random sampling means that each individual is chosen at random from a known location, e.g., the individuals are sampled without taking into account their genotype, ancestry or any other properties apart from location.

- In a population with $D$ demes, there are $\frac{1}{2} D(D + 1)$ possible configurations of a sample of size 2:

$$\frac{D(D + 1)}{2} = \begin{cases} D & \text{same location} \\ \binom{D}{2} & \text{different locations} \end{cases}$$

Notice that the order of sampling doesn’t matter here.
Backward migration rates

The migration rates that will appear in the definition of the structured coalescent are different from those that appear in the model describing the evolution of the population forward-in-time. Specifically,

The backwards migration rate $m_{ij}$ is equal to the probability that an adult chosen at random from deme $i$ was born in deme $j$.

The relationship between the forward and backward migration rates depends on the details of the model. For Notohara’s model, when $N$ is large and $q_{ij}$ is small, this relationship is simple:

$$m_{ij} = \frac{N_j q_{ji}}{N_i}.$$
It will also be convenient to define a total backward migration rate.

**Total backward migration rate**

We will let $m_i$ denote the total backward migration rate for a lineage in deme $i$:

$$m_i = \sum_{j \neq i} m_{ij}$$

Notice that $1 - m_i$ is the probably that a lineage in deme $i$ does not migrate in one generation.
Two individuals sampled from different demes

Suppose that our sample contains an individual from deme $i$ and another from deme $j \neq i$. There are four possible events:

- With probability $m_{ik} \cdot (1 - m_j)$, the lineage in deme $i$ moves to deme $k$, while the lineage in deme $j$ remains there.
- Similarly, with probability $m_{jk} \cdot (1 - m_i)$, the lineage in deme $j$ moves to deme $k$, while the lineage in deme $i$ remains there.
- With probability $m_{ik} \cdot m_{jl}$, the lineage in deme $i$ moves to deme $k$ while the lineage in deme $j$ moves to deme $l$. If $l = k$, then the two lineages end up in the same deme.
- With probability $(1 - m_i) \cdot (1 - m_j)$ both lineages remain where they are.

**Remark:** Because we assume that reproduction and population regulation follow migration, lineages that occupy different demes cannot coalesce in that generation.
Two individuals sampled from the same deme

Now suppose that our sample contains two individuals from deme $i$. In this case, there are six kinds of events:

- With probability $(1 - 1/N_i)(1 - m_i)^2$, the two individuals will have different parents, both from deme $i$.
- With probability $2(1 - 1/N_i)(1 - m_i)m_{ij}$, the two individuals will have different parents, one from deme $i$ and the other from deme $j$.
- With probability $(1 - 1/N_i)m_{ij}^2$, the two individuals will have different parents which are both from deme $j$.
- With probability $2(1 - 1/N_i)m_{ij}m_{ik}$, the two individuals will have different parents which are from demes $j$ and $k$.
- With probability $\frac{1}{N_i}(1 - m_i)$, the two individuals have the same parent, which is from deme $i$.
- With probability $\frac{1}{N_i}m_{ij}$, the two individuals have the same parent, which is from deme $j$. 
Scaling Assumptions

The structured coalescent is greatly simplified when the population sizes of the demes are large (say $N_i > 100$) and the backward migration rates are small. To this end, we will make the following scaling assumptions:

1. The population size of deme $i$ will be denoted $N_i = c_i N$, where $N$ is the total population size. We will assume that $N$ is large and that none of the $c_i$'s are very close to zero.

2. We will assume that each of the backward migration rates is of order $O(1/N)$, i.e.,

$$m_{ij} = \frac{1}{2N} M_{ij},$$

where the parameter $M_{ij}$ is the scaled backwards migration rate.

Rationale: These two assumptions guarantee that coalescence and migration both occur on the same time scale, i.e., it takes on average $O(N)$ generations for either a coalescence or a migration event to occur in the ancestral process.
Structured coalescent for a sample of size 2

With these scaling assumptions, the genealogy for a sample of size two can be approximated by the following process. Here time is measured in units of $N$ generations and the approximation is exact in the limit as $N \to \infty$.

1. If both lineages occupy deme $i$, then they coalesce at rate $1/c_i$, in which case the genealogy has reached the most recent common ancestor and the process is stopped.

2. If both lineages occupy deme $i$, then at rate $M_{ij}$ one of them, chosen uniformly at random, moves to deme $j$ while the other remains in deme $i$.

3. If the lineages occupy separate demes, say $i$ and $j$, then at rate $M_{ik}$ the lineage in deme $i$ moves to deme $k$ while the lineage in deme $j$ remains put. Similarly, at rate $M_{jk}$, the lineage in deme $j$ moves to deme $k$ while the lineage in deme $i$ remains put.
Structured coalescent for a sample of arbitrary size

To describe the structured coalescent for samples containing more than two individuals, let $n_i(t) = n_i$ denote the number of lineages ancestral to the sample that are contained in deme $i$ at time $t$. As before, time is measured in units of $N$ generations.

1. At rate $\binom{n_i}{2} \frac{1}{c_i}$, two lineages in deme $i$ are chosen at random and coalesce, reducing the number of lineages in this deme to $n_i - 1$.

2. At rate $n_i M_{ij}/2$, a lineage is chosen at random from deme $i$ and is moved to deme $j$, reducing $n_i$ to $n_i - 1$ and increasing $n_j$ to $n_j + 1$.

3. This process continues until the final pair of lineages coalesces, in which case the genealogy has reached the most recent common ancestor of the entire sample.
Like Kingman’s Coalescent, the structured coalescent is a **continuous-time Markov chain**.

### Continuous-time Markov Chains

A stochastic process $X = (X(t) : t \geq 0)$ is said to be a continuous-time Markov chain if we only need to know the present state of the process in order to predict its future states, i.e.,

$$
\mathbb{P}(X(t + s) = e | X(u) : 0 \leq u \leq t) = \mathbb{P}(X(t + s) = e | X(t))
$$

for all times $t, s \geq 0$ and every state $e$.

The **state space** of a Markov chain is the set $E$ of values that the random variables $X_t$ can assume, i.e., the set of possible states of the process.
Any continuous-time Markov chain with state space \( E = \{e_1, \cdots, e_n\} \) can be described by an \( n \times n \) matrix of numbers \( Q = (q_{ij}) \) which we call the \textbf{transition matrix} of the process.

\[
Q = \begin{pmatrix}
q_{11} & q_{12} & \cdots & q_{1n} \\
q_{21} & q_{22} & \cdots & q_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
q_{n1} & q_{n2} & \cdots & q_{nn}
\end{pmatrix}
\]

Here,

- Each entry \( q_{ij} \) (with \( i \neq j \)) is the \textbf{transition rate} from state \( e_i \) to \( e_j \).
- Each entry \( q_{ii} \) on the diagonal is equal to the negative of the \textbf{total transition rate} out of state \( i \), i.e.

\[
q_{ii} = - \sum_{j \neq i} q_{ij}.
\]
The transition matrix determines the behavior of the chain $X$ as follows:

- Suppose that the current state is $X(t) = e_i$ and for each state $e_j$ let $\tau_{ij}$ be an independent exponentially distributed random variable with mean $1/q_{ij}$. We can think of the $\tau_{ij}$ as a collection of random alarm clocks which ring at different rates.

- To determine the next state of the chain, we see which of the random alarm clocks rings first. This is determined by the minimum of the $\tau_{ij}$:

  $$\tau_i = \min\{\tau_{i1}, \ldots, \tau_{i2}\}$$

- Between times $t$ and $t + \tau_{ij}$, the process remains in state $e_i$.

- However, at time $t + \tau_i$, the chain jumps to a new state $e_j$, where $j$ is the index of the clock that rang first: $\tau_i = \tau_{ij}$. 
The Jukes-Cantor substitution model is an example of a continuous-time Markov chain that takes values in the state space $E = \{T, C, A, G\}$. The rate matrix for this chain is simply:

$$Q = \begin{pmatrix}
-\mu & \mu/3 & \mu/3 & \mu/3 \\
\mu/3 & -\mu & \mu/3 & \mu/3 \\
\mu/3 & \mu/3 & -\mu & \mu/3 \\
\mu/3 & \mu/3 & \mu/3 & -\mu
\end{pmatrix}$$

**Interpretation:** Since every off-diagonal element is equal to $\mu/3$, this model makes the assumption that every possible mutation occurs at the same rate, $\mu/3$. Further, the total mutation rate for each base is $\mu$. 
The Gillespie Algorithm

Although the description of the CTMC $X$ given on the previous slide can be used to carry out simulations, there is an alternative construction known as the Gillespie algorithm (Gillespie 1977) which is often more efficient. Suppose that the current state of the chain is $X(t) = e_i$.

- The Gillespie algorithm begins by simulating a single independent exponentially distributed random variable $\tau_i$ with mean $1/|q_{ii}|$. This is the holding time in state $e_i$ and the chain remains in this state between times $t$ and $t + \tau_i$.

- At time $t + \tau_i$ the chain jumps to a new state $e_j$ that is chosen at random from $E$ with a probability proportional to the transition rate $q_{ij}$, i.e., $e_j$ is chosen with probability $q_{ij}/|q_{ii}|$.

The advantage of the Gillespie algorithm is that it reduces the number of random variables that must be simulated per jump from $n - 1$ to 2.
The Gillespie algorithm can be used to simulate the structured coalescent. As before, let \( n_i(t) = n_i \) be the number of lineages ancestral to the sample that are present in deme \( i \) at time \( t \) in the past.

- The total transition rate out of state \((n_1, \cdots, n_D)\) is equal to

\[
\lambda = \sum_{i=1}^{D} \left( \frac{n_i}{2} \right) \frac{1}{c_i} + \sum_{i=1}^{D} \sum_{j \neq i} n_i \frac{M_{ij}}{2}
\]

Thus the holding time in this state (i.e., the time until the next coalescent or migration event) is exponentially distributed with mean \( 1/\lambda \).

- The next event can be determined in three steps: we first choose the location of the event and we then determine the type of event and the lineages affected.
Deme $i$ is chosen with probability proportional to the total rate of the events that involve that deme:

$$
\lambda_i = \binom{n_i}{2} \frac{1}{c_i} + \sum_{j \neq i} n_i \frac{M_{ij}}{2}.
$$

Given that deme $i$ was chosen, two lineages chosen at random from within that deme coalesce with probability equal to $\left(\frac{n_i}{2}\right)/c_i/\lambda_i$.

Otherwise, a single lineage is chosen at random from deme $i$ and moved to a second deme $j$, which is chosen with probability equal to $M_{ij}/M_i$. 
Analytical Results

The complexity of the structured coalescent is such that there are relatively few analytical results. However, one special case that has been studied in detail is the finite version of Wright’s island model. Here we will make the following assumptions:

- The population is subdivided into $D$ demes, each containing $N$ haploid individuals.
- Every pair of demes exchanges migrants at the same rate $m/(D - 1)$.

Suppose that we sample two individuals at random from this population. Because of the symmetry of the model, the pairwise coalescent time does not depend on which demes are sampled, only on whether the individuals come from the same deme or from different demes. To this end, let us define

$$T_w = \text{pairwise coalescent time of a sample from the same deme}$$
$$T_b = \text{pairwise coalescent time of a sample from two different demes}$$
Using the Markov property of the structured coalescent, it can be shown that the expected values of $T_w$ and $T_b$ (in units of $N$ generations) satisfy the following recursive equations:

\[
\mathbb{E}[T_w] = \frac{1}{1 + M} + \left( \frac{M}{1 + M} \right) \cdot \mathbb{E}[T_b]
\]

\[
\mathbb{E}[T_b] = \frac{1}{M} + \left( \frac{1}{D - 1} \right) \cdot \mathbb{E}[T_w] + \left( \frac{D - 2}{D - 1} \right) \cdot \mathbb{E}[T_b].
\]

Here $M = 2Nm$ is the total scaled rate of migration. This is a system of two linear equations in two unknowns which can be solved to find

### Mean pairwise coalescent times in the finite island model

\[
\mathbb{E}[T_w] = D
\]

\[
\mathbb{E}[T_b] = D + \frac{D - 1}{M}.
\]
A similar calculation gives:

Variances of the pairwise coalescent times in the finite island model

\[
\begin{align*}
\text{Var}(T_w) &= D^2 + 2 \frac{(D - 1)^2}{M} \\
\text{Var}(T_b) &= D^2 + 2 \frac{(D - 1)^2}{M} + \frac{(D - 1)^2}{M^2}.
\end{align*}
\]

The figure shows how the means (left) and variances (right) of the two pairwise coalescent times vary with the scaled total migration rate \(M\) for an island model with \(D = 10\) demes.
In the strong migration limit, we keep the migration rates fixed and let the deme sizes tend to infinity. Under this scaling, migration occurs much more rapidly than coalescence and it is known that the genealogy can be approximated by Kingman’s coalescent (Notohara 1993).

For example, if the population is subdivided into two demes with sizes $N_1 = c_1 N$ and $N_2 = c_2 N$ and backward migration rates $m_{12}$ and $m_{21}$, then for large $N$, the population is essentially panmictic and the effective population size is a weighted harmonic mean of $N_1$ and $N_2$:

$$N_e = \frac{1}{\frac{m_{21}^2}{(m_{12}+m_{21})^2} \frac{1}{N_1} + \frac{m_{12}^2}{(m_{12}+m_{21})^2} \frac{1}{N_2}}$$
Bayesian Inference of Population Structure

The structured coalescent can be used to carry out Bayesian analyses of sequence data sampled from subdivided populations.

- The objective is to sample from the posterior distribution of the deme sizes and backward migration rates:

  \[ p(\mathcal{G}, (N_i, m_{ij}), \theta_{\text{subst}}|D) \propto p(\mathcal{G}|(N_i, m_{ij})) \cdot p(D|\mathcal{G}, \theta_{\text{subst}}) \]

- In this setting, the genealogy \( \mathcal{G} \) includes the ancestral locations of each branch as well as the times when migration events occur.

- The prior distribution \( p(\mathcal{G}|(N_i, m_{ij})) \) on the unknown genealogy is supplied by the structured coalescent.

- This is often a computationally difficult problem: not only are there many parameters to be estimated, but the state space is very large.

- Implementations include \textbf{Migrate} (Beerli 2006) and \textbf{Lamarc} (Kuhner & Smith 2007).
In view of the difficulty of estimation using the structured coalescent, Lemey et al. (2009) proposed an alternative model (‘Bayesian phylogeographical model’) for analyzing sequence data from a subdivided population.

- In this approach, the location of each ancestral lineages is treated as a discrete neutral character with states \( \{1, \cdots, D\} \).
- Evolution of this character is governed by a continuous-time Markov chain with unknown rate matrix \( \Lambda = (\lambda_{ij}) \). Loosely speaking, \( \lambda_{ij} \) is a proxy for the forward migration rate from location \( i \) to location \( j \).
- The objective is to estimate the posterior distribution of both \( \Lambda \) and the unknown genealogy \( G \), including the locations of the ancestral lineages.
- This approach does not explicitly account for population structure: the prior distribution of the unlabeled genealogy \( G \) is given by an unstructured coalescent. Furthermore, this model does not allow the coalescent rate to vary between locations, although it can vary across time.
Phylogeography of Rabies in African Dogs

From Lemey et al. (2009), using the discrete phylogeographical model.
The version of the Bayesian phylogeographical model implemented in BEAST assumes that the rate matrix $\Lambda$ is of the form

$$\Lambda = \mu S \Pi$$

where

- $\mu$ is the expected number of transitions (‘migrations’) per unit time.
- $S = (s_{ij})$ is a symmetric matrix ($s_{ij} = s_{ji}$) and $s_{ij}$ is the ‘relative migration rate’ from location $i$ to location $j$.
- $\Pi = \text{diag}(\pi_1, \cdots, \pi_D)$ is a diagonal matrix with entries $\pi_1 = \cdots = \pi_D = \frac{1}{D}$.

These assumptions guarantee that the migration process is time-reversible and that at equilibrium a lineage is equally likely to occupy any one of the $D$ locations. While these conditions will usually be violated in real populations, they reduce the number of parameters that have to be estimated.
Bayesian Stochastic Search Variable Selection

One of the strengths of the Bayesian phylogeographical model implemented in BEAST is that it uses a model selection strategy known as Stochastic Search Variable Selection (SSVS) to find a parsimonious migration rate matrix.

- SSVS was originally introduced in the context of Bayesian regression, where the objective was to decide which variables to include in a linear regression model.
- In the phylogeographical context, the objective is to decide which transition rates $s_{ij}$ should be positive.
- SSVS is implemented by multiplying each rate $s_{ij}$ by an indicator variable $\delta_{ij}$ which can either be equal to 0 or 1. If $\delta_{ij} = 0$, then there is no direct migration between locations $i$ and $j$.
- The prior distribution on the number of non-zero indicator variables is taken to be a Poisson distribution with mean $\eta = \ln(2)$ conditioned to be greater than or equal to $D - 1$.
- The lower bound is necessary because there must be at least $D - 1$ positive migration rates for the population to be connected.
Bayesian Phylogeography: Interpreting the Results

The results of an analysis using the Bayesian phylogeographical model include the posterior distribution of the average ‘migration rate’ $\mu$ and of each relative rate $s_{ij}$ and the corresponding indicator variable $\delta_{ij}$.

- The units of $\mu$ will be migrations per unit time, where the latter is determined by the time units of the substitution process (e.g., millions of years).
- The migration rate between locations $i$ and $j$ can be estimated by multiplying the posterior mean of the relative rate $s_{ij}$ by the posterior mean of $\mu$.
- The posterior distribution of $\delta_{ij}$ is determined by the probability that the variable is equal to 1, i.e., $p_{ij} = P(\delta_{ij} = 1)$. This is reported by Tracer.
- Values of $p_{ij}$ close to 1 indicate that the data provides strong support for direct migration between locations $i$ and $j$. 
The level of support for direct migration between locations $i$ and $j$ can be assessed using Bayes factors, which compare the posterior and prior odds of models with and without direct migration between those locations. These are given by the following expression.

**Bayes factors in the phylogeographical model**

$$BF_{ij} = \frac{p_{ij}}{1 - p_{ij}} / \frac{q_{ij}}{1 - q_{ij}}.$$  

Here $q_{ij}$ is the prior probability of direct migration between $i$ and $j$. In the current implementation, each of the $\binom{D}{2}$ rates $s_{ij}$ is equally likely to be positive and so

$$q_{ij} = \frac{\eta + D - 1}{D(D - 1)/2}.$$

**Interpretation of Bayes factors**

As a rule of thumb, rates with Bayes factors greater than 3 can be considered to be well supported by the data.
It is also possible to extend the structured coalescent to models in which the number of demes and the migration rates between demes are variable. Isolation with migration (IM) models are one example.

The simplest IM model assumes that:

- Up until some time $T_A$, the population was panmictic, of size $N_A$.
- At time $T_A$, the ancestral population split into two demes containing $N_1$ and $N_2$ individuals.
- These demes exchange migrants at rates $m_1$ and $m_2$.

The objective is to estimate the three population sizes, the two migration rates and the split time $T_A$. 

Hey & Nielsen 2004

Jay Taylor (ASU)
To describe the structured coalescent in an IM model we need to specify what happens when populations ‘split’.

- One possibility is to assume an instantaneous split of the ancestral population into the two derived populations, in which case any ancestral lineages extant at time $T_A$ move without coalescing into the ancestral population.

- Alternatively, we could assume that the $i$’th derived population was founded by $N_{i,f}$ individuals, which then gave birth to $N_i$ descendants. If $N_{i,f}$ is small, then the genealogy may contain multiple mergers associated with this bottleneck.

The programs **IM**, **IMa** and **IMa2** can be used to perform Bayesian analysis of multilocus sequence data using an IM model with up to ten subpopulations. These are available at Jody Hey’s website (link).
For many problems in population genetics, the evaluation of the full likelihood function of the sequence data is hard. An alternative approach, known as ABC, is to use a combination of simulations and summary statistics to forgo this step.

To apply ABC to a problem, we first need to choose a set of summary statistics \( \eta_1, \cdots, \eta_m \) that carry information about the processes of interest to us. Examples include:

- the numbers of segregating sites and alleles; nucleotide diversity;
- the number of singleton sites in a sample;
- the number of private alleles in each subpopulation;
- pairwise \( F_{st} \) values.

The success of ABC depends on the information content of these summary statistics.
Having chosen the summary statistics, we then carry out the following steps:

1. We sample a value $\theta$ from the prior distribution $p(\theta)$ of the unknown parameters.
2. We simulate the model assuming that $\Theta = \theta$ and we determine the value of the summary statistics $\tilde{\eta}_i$ for that simulation.
3. We compare the simulated and observed summary statistics. If they are sufficiently close, e.g., if $|\eta_i - \tilde{\eta}_i| < \epsilon$, for each $i$, we accept that proposed parameter value $\theta$. Otherwise, we reject it.
4. We repeat steps 1 - 3 until we have accepted a sufficient number of parameter values, say $\theta^{(1)}, \ldots, \theta^{(N)}$.

The accepted parameter values can be used to approximate the posterior distribution of $\Theta$ given the summaries of the full data, i.e.,

$$p(\Theta|\eta_1, \cdots, \eta_m) \approx \frac{1}{N} \sum_{j=1}^{N} \delta_{\theta^{(j)}}$$
Strengths and Weaknesses of ABC

- ABC algorithms are relatively easy to program, making it possible to perform inference for models that aren’t implemented in existing software.
- ABC algorithms may converge when full-likelihood based approaches do not.
- On the other hand, because ABC is based on summary statistics, it does not make full use of the data, i.e., some information is being neglected.
- As the number of summary statistics increases, the probability of accepting a proposed parameter value decreases and so the implementation and execution of ABC becomes more demanding.
Multiple independent introductions of Plasmodium falciparum in South America (Yalcindag et al., PNAS 2011)

Independent Introduction Scenario (A)  
Serial Introduction Scenario (B)  
Unsampled population Scenario (C)
Structured coalescents can also be used to describe the genealogy of a random sample from a population under selection. Here the challenge is to overcome the lack of exchangeability between individuals with different fitnesses. The following observations explain how this can be done using a structured coalescent.

- We can think of the population as being subdivided into groups of chromosomes that carry the same allele at the selected locus (‘genetic structure’).
- Chromosomes that carry the same selected allele will be exchangeable.

The application of structured coalescents to models with selection was first proposed by Hudson et al. (1988) and formalized by Barton et al. (2004).
Selected and Marker Loci

We will consider a model with two loci: a neutral marker locus which is linked to a second locus segregating two alleles $A_1$ and $A_2$ under selection. Our objective is to describe the genealogy of a random sample at the marker locus. To this end, we make the following assumptions:

1. The population size $N$ is large and constant.
2. The population is governed by the Wright-Fisher model with selection: children choose their parents with probabilities proportional to the fitness of each parent.
3. $A_1$ mutates to $A_2$ at rate $\mu_2/N$, while $A_2$ mutates to $A_1$ at rate $\mu_1/N$.
4. Recombination between the selected locus and the marker locus occurs at rate $\rho/N$.

A sample of size 3.
Under fairly general conditions, the structured coalescent for this model can be represented as a continuous-time Markov process

$$\Gamma(t) = (n_1(t), n_2(t), p(t)),$$

where

- $n_i(t)$ is the number of ancestral lineages that carry the $A_i$ allele.
- $p(t)$ is the frequency of allele $A_1$ at time $t$ in the past.
- We will also use $q(t) = 1 - p(t)$ to denote the frequency of allele $A_2$ at time $t$ in the past.
Changes to $\Gamma(t)$ occur through the following events:

1. Two $A_1$ lineages can coalesce.
2. Two $A_2$ lineages can coalesce.
3. Each lineage can migrate between backgrounds, through:
   - mutation at the selected locus;
   - recombination between the selected and marker loci.
4. The ancestral allele frequencies change through selection, mutation and genetic drift.
Transition Rates of the Structured Coalescent

When $N$ is large and time is measured in units of $N$ generations, the transition rates for this process are given by the following expressions:

<table>
<thead>
<tr>
<th>Transition</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>two $A_1$ lineages coalesce</td>
<td>$(\frac{n_1}{2}) \frac{1}{p}$</td>
</tr>
<tr>
<td>two $A_2$ lineages coalesce</td>
<td>$(\frac{n_2}{2}) \frac{1}{q}$</td>
</tr>
<tr>
<td>a lineage mutates from $A_1$ to $A_2$</td>
<td>$n_1 \mu_1 (q/p)$</td>
</tr>
<tr>
<td>a lineage mutates from $A_2$ to $A_1$</td>
<td>$n_2 \mu_2 (p/q)$</td>
</tr>
<tr>
<td>a lineage recombines from $A_1$ to $A_2$</td>
<td>$n_1 rq$</td>
</tr>
<tr>
<td>a lineage recombines from $A_2$ to $A_1$</td>
<td>$n_2 rp$</td>
</tr>
</tbody>
</table>

These are analogous to the rates that we derived for geographically structured models. For example, the coalescent rate of ancestral lineages carrying the $A_1$ allele is inversely proportional to the total number of $A_1$ alleles in the population, i.e., $Np$. 
Diffusion Approximations in Population Genetics

To complete the description of the structured coalescent we need to specify how the frequencies of the selected alleles change over time. This is usually done with a diffusion approximation.

**Diffusion Processes**

A diffusion process is a continuous-time Markov process \( (X(t) : t \geq 0) \) with continuous sample paths.

- Diffusion approximations are accurate when \( N \) is large and selection and mutation act on coalescent time scales.
- Stationary diffusion processes have the convenient property of being reversible: we recover the same diffusion process under time reversal.
Taylor (2013) used a structured coalescent to study the effects of fluctuating selection on genealogies at linked sites.

- In this model, the fitnesses of the alleles $A_1$ and $A_2$ fluctuate randomly from one generation to the next.
- Fluctuating selection accelerates coalescence at the selected locus, resulting in shallower genealogies.
- However, fluctuating selection can also increase the heterozygosity at the selected locus.
Contrasting Balancing and Fluctuating Selection

Although balancing selection and fluctuating selection can both maintain excess polymorphism at the selected locus, they have different effects on linked neutral sites.

Balancing Selection:
- Coalescent times and heterozygosities are elevated at tightly linked sites, e.g., near the ADH locus in *D. melanogaster*.

Fluctuating Selection:
- Coalescent times and heterozygosities are reduced over an extensive range of linked sites.